IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 79–94
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 17/06/13

Thematically Reinforced Explicit Semantic Analysis

YANNIS HARALAMBOUS[1] AND VITALY KLYUEV[2]

[1] *Institut Mines-Télécom - Télécom Bretagne and Lab-STICC UMR CNRS 6285, France*
[2] *University of Aizu, Japan*

ABSTRACT

We present an extended, thematically reinforced version of Gabrilovich and Markovitch's Explicit Semantic Analysis (ESA), where we obtain thematic information through the category structure of Wikipedia. For this we first define a notion of categorical tfidf which measures the relevance of terms in categories. Using this measure as a weight we calculate a maximal spanning tree of the Wikipedia corpus considered as a directed graph of pages and categories. This tree provides us with a unique path of "most related categories" between each page and the top of the hierarchy. We reinforce tfidf of words in a page by aggregating it with categorical tfidfs of the nodes of these paths, and define a thematically reinforced ESA semantic relatedness measure *which is more robust than standard ESA and less sensitive to noise caused by out-of-context words. We apply our method to the French Wikipedia corpus, evaluate it through a text classification on a 37.5 MB corpus of 20 French newsgroups and obtain a precision increase of 9–10% compared with standard ESA.*

1 INTRODUCTION

1.1 *Explicit Semantic Analysis*

Unlike semantic similarity measures, which are limited to ontological relations such as synonymy, hyponymy, meronymy, etc., *semantic relatedness*

measures detect and quantify semantic relations of a more general kind. The typical example is the one involving the concepts CAR, VEHICLE and GASOLINE. A car is a special kind of vehicle, so we have an hyperonym relation between the concepts, which can easily be quantified by a semantic similarity measure (for example, by taking the inverse of the length of the shortest path between the corresponding synsets in WordNet). But between CAR and GASOLINE, there is no semantic *similarity*, since a car is a solid object and fuel is a liquid. Nevertheless, there is an obvious semantic relation between them since most cars use gasoline as their energy source, and such a relation can be quantified by a semantic *relatedness* measure.

Gabrilovich & Markovitch [1] introduce the semantic relatedness measure ESA (= Explicit Semantic Analysis, as opposed to the classical method of Latent Semantic Analysis [2]). ESA is based on the Wikipedia corpus. Here is the method: after cleaning and filtering Wikipedia pages (keeping only those with a sufficient amount of text and a given minimal number of incoming and outgoing links), they remove stop words, stem all words and calculate their tfidfs. Wikipedia pages can then be represented as vectors in the space of (nonempty, stemmed, distinct) words, the vector coordinates being normalized tfidf values. By the encyclopedic nature of Wikipedia, one can consider that every page corresponds to a concept. We thus have a matrix whose columns are concepts and whose lines are words. By transposing it we obtain a representation of words in the space of concepts. The ESA measure of two words is simply the cosine of their vectors in this space.

Roughly, two words are closely ESA-related *if they appear frequently in the same Wikipedia pages* (so that their tfs are high), *and rarely in the corpus as a whole* (for their dfs to be low).

Despite the good results obtained by this method, it has given rise to some criticism. Thus, Haralambous & Klyuev [3] note that ESA has poor performance when the relation between words is mainly ontological. As an example, in the English corpus, the word “mile” (length unit) does not appear in the page of the word “kilometer” and the latter appears only once in the page of the former: this is hardly sufficient to establish a nonzero semantic relatedness value; however, such a relation is obvious, since both words refer to units of length measurement. As pointed out in [3], an ontological component, obtained from a WordNet-based measure, can, at least partially, fill this gap.

Another, more fundamental, criticism is that of Gottron et al. [4], who argue that the choice of Wikipedia is irrelevant, and that any corpus of comparable size would give the same results. To prove it, they base ESA not on Wikipedia, but on the Reuters news corpus, and get even better results

than with standard ESA. According to the authors, the semantic relatedness value depends only on the collocational frequency of the terms, and this whether documents correspond to concepts or not. In other words they deny the "concept hypothesis," namely that ESA specifically uses the correspondence between concepts and Wikipedia pages. Also they state that while "the application of ESA in a specific domain benefits from taking an index collection from the same topic domain while, on the other hand, a "general topic corpus" such as Wikipedia introduces noise," and this has precisely been our motivation for strengthening the thematic robustness of ESA. Indeed, in this article we will enhance ESA by adopting a different approach: the persistence of tfidfs of terms when leaving pages and entering the category graph.

1.2 *Wikipedia Categories*

A Wikipedia page can belong to one or more categories. Categories are represented by specific pages using the "Category:" prefix; these pages can again belong to other categories, so that we obtain a directed graph structure, the nodes of which can be standard pages (only outgoing edges) or categories (in- and outgoing edges). A page can belong to several categories and there is no ranking of their semantic relevance. For this reason, to be able to use categories, we first need an algorithm to determine the single semantically most relevant category, and for this we use, once again, ESA.

Wikipedia's category graph has been studied thoroughly in [5] (for the English corpus).

1.3 *Related Work*

Scholl et al. [6] also enhance the performance of ESA using categories. They proceed as follows: let T be the matrix whose rows represent the Wikipedia pages and whose columns represent words. The value $t_{i,j}$ of cell (i, j) is the normalized tfidf of the jth word in the ith page. For each word m there is therefore a vector $\boldsymbol{v}_m$ whose dimension is equal to the number of pages. Now let C be the matrix whose columns are pages and whose lines are categories. The value of a cell $c_{i,j}$ is 1 when page j belongs to category i and 0 otherwise. They take the product of matrices $\boldsymbol{v}_m \cdot C$ which provides a vector whose jth component is $\sum_{i|D_i \in c_j} t_{i,j}$, that is the sum of tfidfs of word m for all pages belonging to the jth category. They use the concatenation of vector $\boldsymbol{v}_m$ and of the transpose of $\boldsymbol{v}_m \cdot C$ to improve

system performance on the text classification task. They call this method XESA (eXtended ESA).

We see that in this attempt, page tfidf is extended to categories by simply taking the sum of tfidfs of all pages belonging to a given category. This approach has a disadvantage when it comes to high-level categories: instead of being a way to find the words that characterize a given category, the tfidf of a word tends to become nothing more than the average density of the word in the corpus, since for large categories, tf tends to be the total number of occurrences of the word in the corpus, while the denominator idf remains constant and equal to the number of documents containing the given word. Thus, this type of tfidf loses its power of discrimination for high-level categories. As we will see in Section 2.2, we propose another extension of tfidf to categories, which we call *categorical tfidf*. The difference lies in the denominator, where we take the number, not of all documents containing the term, but only of those *not belonging* to the category. Thus our categorical tfidf (which is equal to the usual tfidf in the case of pages) is high when the term is common in the category and *rare elsewhere* (as opposed to *rare on the entire corpus* of Scholl et al.).

In [7], the authors examine the problem of inconsistency of Wikipedia's category graph and propose a shortest path approach (based on the number of edges) between a page and the category "*Article*," which is at the top of the hierarchy. The shortest path provides them with a semantic and thematic hierarchy and they calculate similarity as shortest length between vertices on these paths, a technique already used in WordNet [8]. However, as observed in [8, p. 275], the length (in number of edges) of the shortest path can vary randomly, depending on the density of pages (synsets, in the case of WordNet) in a given domain of knowledge. On the other hand, the distance (in number of edges) between a leaf and the top of the hierarchy is often quite short, frequently requiring an arbitrary choice between paths of equal length.

What is common with our approach is the intention to simplify Wikipedia's category graph. But instead of counting edges, we weight the graph using ESA measure and use this weight, which is based on the statistical presence of words on pages belonging to a given category, to calculate a maximum spanning tree. The result of this operation is that any page (or category other than "*Article*") has exactly one parent category that is semantically closest to it. This calculation is global, in the sense that the total weight of the tree is maximum.

We use this tree to define *thematically reinforced ESA*. Our goal is to avoid words which, by accident, have a high tfidf in a given page despite the fact that they thematically do not really belong to it. This happens in

the very frequent case where words have low frequencies (in the order of 1–3) so that the presence of an unsuitable word in a page results in a tfidf value as high (or even higher, if the word is seldom elsewhere) as the one of relevant words. Our hypothesis is that a word having an unduly high tfidf will disappear when we calculate its (categorical) tfidf in categories above the page, while, on the contrary, relevant words will be shared by other pages under the same category and their tfidfs will continue to be nonzero when switching to them. Such words will "survive" when we move away from leaves of the page-and-category tree and towards the root.

2 THEMATIC REINFORCEMENT

2.1 *Standard Tfidf, Concept Vector and ESA Measure*

Let us first formalize the standard ESA model.[3]

Let $\mathcal{W}$ be the Wikipedia corpus pruned by the standard ESA method, $p \in \mathcal{W}$ a Wikipedia page, and $w \in p$ a word.[4] The tfidf $t_p(w)$ of the word w on page p is defined as:

$$t_p(w) := (1 + \log(f_p(w))) \cdot \log\left(\frac{\#\mathcal{W}}{\sum_{\substack{p\in\mathcal{W}\\ w\in p}} 1}\right),$$

where $f_p(w)$ is the frequency of w on page p, $\#\mathcal{W}$ the cardinal of $\mathcal{W}$ and $\sum_{\substack{p\in\mathcal{W}\\ w\in p}} 1$, also known as the df (= document frequency) of w, is the number of Wikipedia pages containing w.

Consider the space $\mathbb{R}^{\#\mathcal{W}}$, where dimensions correspond to pages p of $\mathcal{W}$. Then we define the "concept vector" $\boldsymbol{w}$ of word w as

$$\boldsymbol{w} := \sum_{p\in\mathcal{W}} t_p(w) \cdot 1_p \in \mathbb{R}^{\#\mathcal{W}}$$

where 1_p is the unitary vector of $\mathbb{R}^{\#\mathcal{W}}$ corresponding to page p.

Let w and w' be words appearing in Wikipedia (and hence the Euclidean norms $\|\boldsymbol{w}\|$ and $\|\boldsymbol{w'}\|$ of their concept vectors are nonzero). The ESA semantic relatedness measure μ is defined as follows:

$$\mu(w, w') := \frac{\langle \boldsymbol{w}, \boldsymbol{w'}\rangle}{\|\boldsymbol{w}\| \cdot \|\boldsymbol{w'}\|}.$$

[3] All definitions in Section 2.1 are from [1].

[4] By "word" we mean an element of the set of character strings remaining after removing stopwords and stemming the Wikipedia corpus.

2.2 *Categorical Tfidf*

Let c be a Wikipedia category. We define $\mathcal{F}(c)$ as the set of all pages p such that

- either p belongs to c,
- or p belongs to c_1, and there a sequence of subcategory relations $c_1 \to c_2 \to \cdots \to c$, ending with c.

Definition 1 *Let $w \in p$ be a word of $p \in \mathcal{W}$, $t_p(w)$ its standard tfidf in p, and c a category of $\mathcal{W}$. We define the* categorical tfidf $t_c(w)$ *of w for category c as follows:*

$$t_c(w) := \left(1 + \log\left(\sum_{p \in \mathcal{F}(c)} f_p(w)\right)\right) \cdot \left(\log\left(\frac{\#\mathcal{W}}{1 + \sum_{\substack{p \in \mathcal{W} \setminus \mathcal{F}(c) \\ w \in p}} 1}\right)\right).$$

The difference with the tfidf defined by [6] is in the calculation of df: instead of $\sum_{\substack{p \in \mathcal{W} \\ w \in p}} 1$, that is the amount of pages containing w in the entire Wikipedia corpus, we focus on those in $\mathcal{W} \setminus \mathcal{F}(c)$, namely the set difference between the whole corpus and pages that are ancestors of c in the category graph, and we use $1 + \sum_{\substack{p \in \mathcal{W} \setminus \mathcal{F}(c) \\ w \in p}} 1$ instead (the unit is added to prevent a zero df in the case where the word does not appear outside $\mathcal{F}(c)$). We believe that this extension of tfidf to categories improves discriminatory potential, even when the sets of pages become large (see discussion in Section 1.3).

2.3 *Vectors of Pages and Categories*

Let $p \in \mathcal{W}$ be a page. We define the *page vector* $\boldsymbol{p}$ as the normalized sum of concept vectors of its words, weighted by their tfidfs:

$$\boldsymbol{d} := \frac{\sum_{w \in p} t_p(w) \cdot \boldsymbol{w}}{\| \sum_{w \in p} t_p(w) \cdot \boldsymbol{w} \|}.$$

Similarly let c be a category of Wikipedia, we define the *category vector* $\boldsymbol{c}$ as

$$\boldsymbol{c} := \frac{\sum_{w \in \mathcal{F}(c)} t_c(w) \cdot \boldsymbol{w}}{\| \sum_{w \in \mathcal{F}(c)} t_c(w) \cdot \boldsymbol{w} \|}.$$

where $w \in \mathcal{F}(c)$ means that there exists a page p such that $p \in \mathcal{F}(c)$ and $w \in p$.

2.4 *Wikipedia Arborification*

Definition 2 *Let p be a Wikipedia page and c, c' Wikipedia categories. Let $p \to c$ be the membership of page d to category c, and $c \to c'$ the subcategory relation between c and c. We define the* weight of semantic relatedness *of these relations as*

$$p(p \to c) = \langle \boldsymbol{p}, \boldsymbol{c} \rangle.$$
$$p(c \to c') = \langle \boldsymbol{c}, \boldsymbol{c'} \rangle,$$

where $\langle \, . \, , \, . \, \rangle$ is the Euclidean scalar product of two vectors.

This product is equal to the cosine metric since the vectors are all unitary. By this property we also have $\mathrm{Im}(p) \subset [0, 1]$.

The relations considered in Definition 2 correspond to vertices of the Wikipedia category graph. Let $\mathcal{W}'$ be the weighted Wikipedia digraph; its vertices are pages and categories, its edges are memberships of pages and inclusions of categories, and its weight is the weight of semantic relatedness.

At this point we can already reinforce the standard tfidf of words on pages, by the categorical tfidf of the same words in related categories. But how can we choose these categories? Taking all those containing a page would result in cacophony since categories can be more or less relevant and sometimes have no semantic relation whatsoever. Not to mention the fact that the Wikipedia category graph is quite complex, and using it as such would be computationally prohibiting.

The solution we present to this problem is to simplify $\mathcal{W}'$ by extracting a maximal spanning tree. It should be noted that standard minimal/maximal spanning tree algorithms such as Kruskal or Prim cannot be applied because $\mathcal{W}'$ is directed, has a global sink, namely the "Article" page, and we want the orientation of the directed spanning tree to be compatible with the one of the directed graph[5].

To obtain the maximal spanning tree, we utilized Chu-Liu & Edmonds' algorithm [9, p. 113-119], published for the first time in 1965. This semi-linear algorithm returns a minimum weight forest of rooted trees covering the digraph. The orientation of these rooted trees is compatible with the one of the graph. In the general case, connectivity is not guaranteed (even though the graph may be connected). But in the case of a digraph containing a global sink, the forest becomes a single tree, and we get a true *directed*

[5] It is a known fact that every rooted tree has exactly two possible orientations: one going from the root to the leaves and one in the opposite direction.

maximal spanning tree of the graph. If our case, the global sink is obviously the category that is hierarchically at the top, namely "*Article*."[6]

Let $\mathcal{T}$ be the maximal spanning tree of $\mathcal{W}'$ obtained by our method. As in any tree, there is a unique path between any two nodes. In particular, there is a unique path between any page-node and the root; we call it the *sequence of ancestors* of the page.

2.5 *Thematically Reinforced ESA*

We will use the page ancestors in the maximal spanning tree to update tfidf values of words in the page vectors. Indeed, a word in a given page may have a high tfidf value simply because it occurred one or two times, this does not guarantee a significant semantic proximity between the word and the page. But if the word appears also in ancestor categories (and hence, in other pages belonging to the same category), then we have stronger chances for semantic pertinence.

Definition 3 *Let p be a Wikipedia page, w a word $w \in p$, $t_p(w)$ the standard tfidf of w in p, $(\pi^i(p))_i$ the sequence of ancestors of p, and $(\lambda_i)_i$ a decreasing sequence of positive real numbers converging to 0. We define the* thematically reinforced tfidf $t_{p,\lambda_*}(w)$ *as*

$$t_{p,\lambda_*}(w) = t_p(w) + \sum_{i \geq 0} \lambda_i t_{\pi^i(p)}(w).$$

The sum is finite because the Wikipedia maximal spanning tree is finite and hence there is a maximal distance from the root, after which the π^i become vacuous.

Definition 4 *With the notations of Definition 3, we define the* thematically reinforced concept vector $\boldsymbol{w}_{\lambda_*}$ *as*

$$\boldsymbol{w}_{\lambda_*} := \sum_{p \in \mathcal{W}} t_{p,\lambda_*}(w) \cdot 1_p \in \mathbb{R}^{\#\mathcal{W}}.$$

[6] It should be noted, however, that the path between a page and the root on the maximal spanning tree is not a maximal path per se, since the importance is given to the global maximality of weight, for the whole tree. If our goal were to find the most appropriate taxonomy for a specific page, i.e., the most relevant path from this page to the top, then it would be more appropriate to use a shortest/longest path algorithm, such as Dijkstra. This has already been proposed in [7], but for the metric of the number of edges; in our case we would rather use the measure given by the weight of the graph.

In other words, it is the usual concept vector definition, but using thematically reinforced tfidf.

With these tools we can define our extended version of ESA, as follows:

Definition 5 *With the notations of Definition 3 and $w, w' \in \mathcal{W}$, we define the* thematically reinforced ESA semantic relatedness measure μ_{λ_*} *as:*

$$\mu_{\lambda_*}(w, w') := \frac{\langle \boldsymbol{w}_{\lambda_*}, \boldsymbol{w}'_{\lambda_*} \rangle}{\|\boldsymbol{w}_{\lambda_*}\| \cdot \|\boldsymbol{w}'_{\lambda_*}\|}.$$

In other words, it is the usual ESA measure definition, but using thematically reinforced concept vectors and tfidf.

3 CORPUS

We have chosen to work on the French Wikipedia corpus (version of December 31, 2011), which is smaller than the English one and, to our knowledge, has not yet been used for ESA. To adapt ESA to French Wikipedia, we followed the same steps as [1] and [10] except for one: we have preceded stemming by lemmatization, to avoid loss of information due to poor stemming of inflected words. (In English, inflection is negligible, so that stemming can be performed directly.)

Originally, the authors of [1] pruned the 2005 English Wikipedia corpus down to 132,689 pages. In our case, by limiting the minimum size of pages to 125 (nonstop, lemmatized, stemmed and distinct) words, 15 incoming and 15 outgoing links, we obtained a number of Wikipedia pages comparable to that of the original ESA implementation, namely 128,701 pages (out of 2,782,242 in total) containing 1,446,559 distinct words (only 339,679 of which appear more than three times in the corpus).

Furthermore, the French corpus contains 293,244 categories, 680,912 edges between categories and 12,935,688 edges between pages and categories. As can be seen in Fig. 1, by the logarithmic distribution of incoming and outgoing degrees, this graph follows a power distribution $p^{-\alpha}$ with $\alpha = 2.08$ for incoming degrees and $\alpha = 7.51$ for outgoing degrees. According to [11, p. 248], the former value is typical, while the latter can be considered very high, and this was another motivation for simplifying the Wikipedia graph by extracting the maximal spanning tree, instead of performing heavy calculations on the entire graph.

The French Wikipedia category graph is fairly complex and, in particular, contains cycles. Indeed, according to [12], "cycles are not encouraged

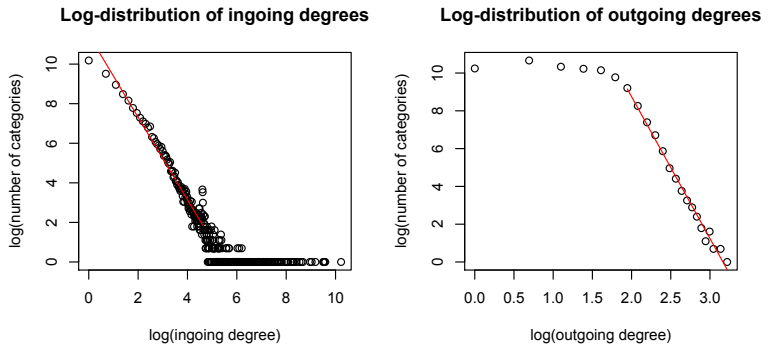

Fig. 1. Ingoing and outgoing degree distribution of the French Wikipedia categories.

but may be tolerated in rare cases." The very simple example of categories "*Zoologie*" (= Zoology) and "*Animal*" (in French Wikipedia) pointing to each other, shows that the semantic relation underlying subcategories is not always hyperonymy. Here ANIMAL is the object of study of the discipline ZOOLOGY. We attempted the following experiment: starting from the 2,782,242 (unfiltered) French Wikipedia pages, we followed random paths formed by the category links. The choice of each subsequent category was made at random, but did not change during the experiment. 78% of these paths contained cycles, but it turned out that it was always the same 50 cycles, 12 of which were of length 3 (triangles) and all others of length 2 (categories pointing to each other, as in the example above, which was detected by this method). Hence, we were able to turn this directed graph acyclic by merely removing 50 edges.

4 EVALUATION

Gabrilovich and Markovitch [1] evaluate their method on WS-353, a set of 352 English word pairs, the semantic relatedness of which has been evaluated by 15–16 human judges. Their criterion is the Spearman correlation coefficient between the rank of pairs obtained by ESA and that obtained by taking the average of human judgments. Our first attempt was to translate these pairs into French, but the result was rather disappointing.[7]

[7] Indeed, some twenty words are untranslatable into a simple term (the current version of ESA covers only single-word terms), such as "seafood" which

We have therefore chosen to evaluate our implementation of ESA in a more traditional way, by performing a text classification task. We have extracted a total of 20,000 French language messages from the 20 most popular French newsgroups. The characteristics of our evaluation corpus can be seen on Table 1, where the second column represents the number of messages for a given newsgroup, the third the number of words, and the fourth, the number of distinct stemmed nonstop words that also occur in Wikipedia.

Table 1. Characteristics of the evaluation corpus

Theme	Newsgroup	# mess.	# words	# terms
Medicine	fr.bio.medecine	1,000	738,258	14.785
Writing	fr.lettres.ecriture	1,000	688,849	14,948
French language	fr.lettres.langue.francaise	1,000	594,143	14,956
Animals	fr.rec.animaux	1,000	391,270	10,726
Classical music	fr.rec.arts.musique.classique	1,000	379,794	15,056
Rock music	fr.rec.arts.musique.rock	1,000	318,434	12,764
Do-it-yourself	fr.rec.bricolage	1,000	358,220	8,349
Movies	fr.rec.cinema.discussion	1,000	680,480	18,284
Gardening	fr.rec.jardinage	1,000	495,465	12,042
Photography	fr.rec.photo	1,000	415,767	10,931
Diving	fr.rec.plongee	1,000	485,059	11,326
Soccer	fr.rec.sport.football	1,000	612,842	13,548
Astronomy	fr.sci.astronomie	1,000	444,576	10,781
Physics	fr.sci.physique	1,000	598,079	13,916
Economics	fr.soc.economie	1,000	737,795	14,797
Environment	fr.soc.environnement	1,000	683,806	15,756
Feminism	fr.soc.feminisme	1,000	612,844	16,716
History	fr.soc.histoire	1,000	675,957	16,458
Religion	fr.soc.religion	1,000	763,477	16,124
Sects	fr.soc.sectes	1,000	738,327	16,732
Global		20,000	11,413,442	67,902

can be translated only as "*fruits de mer*." Furthermore there are ambiguities of translation resulting from word polysemy: When we translate the pair "flight/car" by "*vol*/*voiture*," we obtain a high semantic relatedness due to the criminal sense of "*vol*" (= theft) while the sense of the English word "flight" is mainly confined to the domain of aviation. Finally, some obvious collocations disappear when translating word for word, such as "soap/opera" which is unfortunately not comparable to "*savon*/*opéra*"...

Table 2. Evaluation results (ordered by decreasing precision)

λ_1	λ_2	λ_3	λ_4	λ_5	C	# SVs	Precision	λ_1	λ_2	λ_3	λ_4	λ_5	C	# SVs	Precision
1.5	0	0.5	0.25	0.125	3.0	786	**75.015%**	0	1	0.5	0.25	0.125	3.0	710	74.716%
1	0	0.5	0.25	0.125	3.0	709	74.978%	2	1	0.5	0.25	0.125	3.0	899	74.705%
1.5	1	0.5	0.25	0.125	3.0	827	74.899%	2	0	0.5	0.25	0.125	3.0	852	74.675%
0.25	1.5	0.5	0.25	0.125	3.0	761	74.87%	0.5	0.25	0.125	0.0625	0.0312	3.0	653	74.67%
0.5	0	0.5	0.25	0.125	3.0	698	74.867%	2	0.5	0.5	0.25	0.125	3.0	899	74.641%
1	0.5	0.25	0.125	0.0625	3.0	736	74.845%	0.25	0.125	0.0625	0.0312	0.015	3.0	615	74.613%
0.5	1	0.5	0.25	0.125	3.0	736	74.795%	1	1	1	0.5	0.25	3.0	796	74.61%
1	1.5	0.5	0.25	0.125	3.0	865	74.791%	0	1.5	1	0.5	0.25	3.0	792	74.548%
0.5	0.5	0.5	0.25	0.125	3.0	682	74.789%	1.5	1.5	1	0.75	0.25	3.0	900	74.471%
0.5	1.5	0.5	0.25	0.125	3.0	778	74.814%	2	1.5	1	0.5	0.25	3.0	**995**	74.36%
1.5	0.5	0.2	0.1	0.05	3.0	775	74.780%	0	0	0	0	0	3.0	324	65.58%

To perform text classification we need to extend the definitions of tfidf and document vector to the evaluation corpus. Let $\mathcal{C}$ be the evaluation corpus and d a document $d \in \mathcal{C}$. We define the tfidf $t_d(w)$ of a word $w \in d$ in $\mathcal{C}$ as

$$t_d(w) := (1 + \log(f_d(w))) \cdot \log\left(\frac{\#\mathcal{C}}{\mathrm{df}(w)}\right),$$

where f_d is the frequency of w in d; $\#\mathcal{C}$ the total number of documents; $\mathrm{df}(w)$ the number of documents in $\mathcal{C}$, containing w.

Furthermore, our ESA implementation provides us with a concept vector $\boldsymbol{w}$ for every word w. We define the *document vector* $\boldsymbol{d}$ as:

$$\boldsymbol{d} := \frac{\sum_{w \in d} t_d(w) \cdot \boldsymbol{w}}{\| \sum_{w \in d} t_d(w) \cdot \boldsymbol{w} \|}.$$

where the denominator is used for normalization.

Using these vectors, text classification becomes standard classification in $\mathbb{R}^{\#\mathcal{W}}$ for the cosine metric. We applied the linear multi-class SVM classifier SVM$^{\text{multiclass}}$ [13] to the set of these vectors and the corresponding document classes, and after a tenfold cross-validation, we obtained an average precision of 65.58% for a C coefficient of 3.0. The classification required 324 support vectors. Admittedly the precision obtained is rather low, which is partly due to the thematic proximity of some classes (like, for example, Religion and Sects, or Writing and French language). However, our goal is not to compare ESA to other classification methods, but to show that our approach improves ESA. So, this result is our starting point and we intend to improve it.

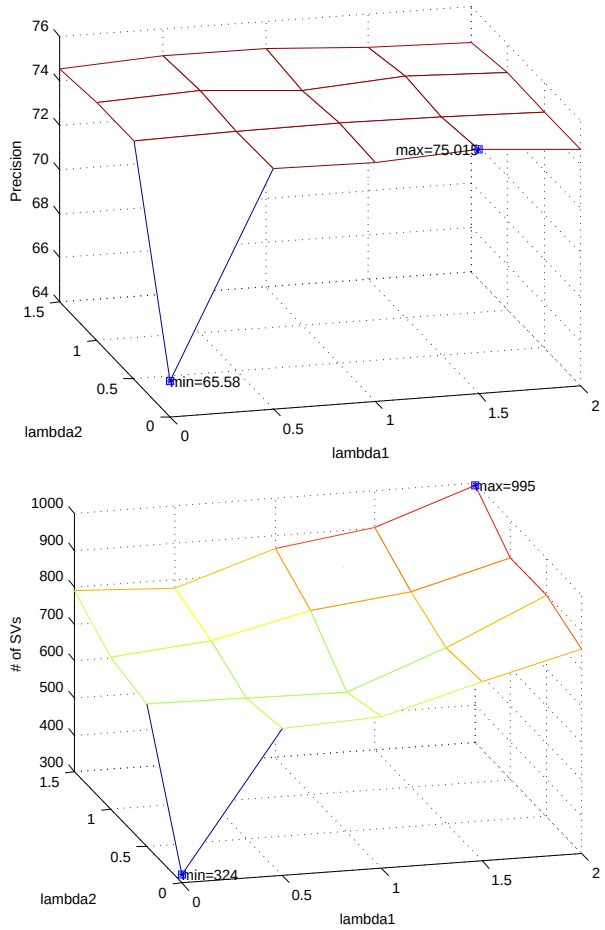

Fig. 2. Precision (to the left) and number of support vectors used (to the right), as functions of the parameters λ_1 and λ_2.

We followed the same modus operandi using thematically reinforced methods and obtained the results displayed on Table 2. The results show a significant improvement over the standard ESA version (that corresponds to $\lambda_i = 0$ for all i. This confirms our approach. In Fig. 2 the reader can see the precision obtained as function of the two first parameters λ_1 and λ_2, as well the number of support vectors used. We notice that the precision

varies slightly (between 74.36% and 75.015%, that is less than 1%) as long as λ_1 or λ_2 are nonzero, and abruptly goes down to 65.58% when they are both zero. For nonzero values of λ_i the variation of precision follows no recognizable pattern. On the other hand, the number of support vectors shows a pattern: it is clearly correlated with λ_1 and λ_2, the highest value being 995, number of support vectors used when both λ_1 and λ_2 take their highest values. Since CPU time is roughly proportional to the number of support vectors, it is most interesting to take small (but nonzero) values of λ_i so that, at the same time, precision is high and the number of support vectors (and hence CPU time) is kept small.

5 Conclusion and Hints for Further Research

By reinforcing the thematic context of words in Wikipedia pages, context obtained through the category structure, we claim to be able to improve the performance of the ESA measure.

We evaluated our method on a text classification task based on messages from the 20 most popular French language newsgroups: thematic reinforcement allowed us to improve the classification precision by 9–10%.

Here are some hints for research to be done:

1. propose the notion of the "most relevant category" to Wikipedia users and use their feedback to improve the system;
2. when we take the "most relevant category" for each page, we don't consider by how much it is better than the others. For small differences of semantic relevance weight between categories one could imagine alternative "slightly worse" spanning trees and compare the results;
3. by comparing relevance between alternative "most relevant" categories for the same page one could quantify a "global potential" of the Wikipedia corpus. Compare with Wikipedia corpora in other languages;
4. aggregate the thematically reinforced measure with collocational and ontological components, as in [3];
5. define another measure, based on links between pages (or categories), proportional to the number of links (or link paths) between pages and inversely proportional to the length of these paths. Compare it to ESA (which uses the number of links between pages to filter Wikipedia, but does not include it in semantic relatedness calculations) and thematically reinforced ESA;
6. and, more generally, explore the applications of graph theory to the formidable mathematical-linguistic objects represented by the different graphs extracted from Wikipedia.

REFERENCES

1. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence (2007)
2. Deerwester, S.C., Dumais, S.T., Furnas, G.W., Harshman, R.A., Landauer, T.K., Lochbaum, K.E., Streeter, L.A.: Computer information retrieval using latent semantic structure (1989) US Patent 4,839,853 of June 13, 1989.
3. Haralambous, Y., Klyuev, V.: A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge. In: International Joint Conference on Natural Language Processing, Chiang-Mai, Thailand (2011)
4. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: CIKM'11: Proceedings of the 20th ACM international conference on Information and knowledge management. (2011)
5. Zesch, T., Gurevych, I.: Analysis of the Wikipedia category graph for NLP applications. In: Workshop TextGraphs-2 : Graph-Based Algorithms for Natural Language Processing. (2007) 1–8
6. Scholl, P., Böhnstedt, D., García, R.D., Rensing, C., Steinmetz, R.: Extended explicit semantic analysis for calculating semantic relatedness of web resources. In: EC-TEL'10: Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL: from innovation to learning and practice, Springer (2010)
7. Collin, O., Gaillard, B., Bouraoui, J.L.: Constitution d'une ressource sémantique issue du treillis des catégories de wikipedia. In: TALN 2010. (2010)
8. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed.: WordNet, an electronic lexical database, The MIT Press (1998) 266–283
9. Gabow, H.N., Galil, Z., Spencer, T., Tarjan, R.E.: Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. Combinatorica **6** (1986) 109–122
10. Çallı, Ç.: Improving search result clustering by integrating semantic information from Wikipedia. Master's thesis, Middle East Technical University, Ankara (2010)
11. Newman, M.: Networks. An Introduction. Oxford University Press (2010)
12. Medelyan, O., Legg, C., Milne, D., Witten, I.H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies **67**(9) (2009) 716–754
13. Joachims, T.: Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning, MIT Press (1999)

YANNIS HARALAMBOUS
INSTITUT MINES-TÉLÉCOM - TÉLÉCOM BRETAGNE,
AND LAB-STICC UMR CNRS 6285,
TECHNOPÔLE BREST-IROISE, CS 83818,
29238 BREST CEDEX 3, FRANCE
E-MAIL: <YANNIS.HARALAMBOUS@TELECOM-BRETAGNE.EU>

VITALY KLYUEV
SOFTWARE ENGINEERING LABORATORY,
UNIVERSITY OF AIZU,
AIZU-WAKAMATSU, FUKUSHIMA-KEN 965-8580, JAPAN
E-MAIL: <VKLUEV@U-AIZU.AC.JP>